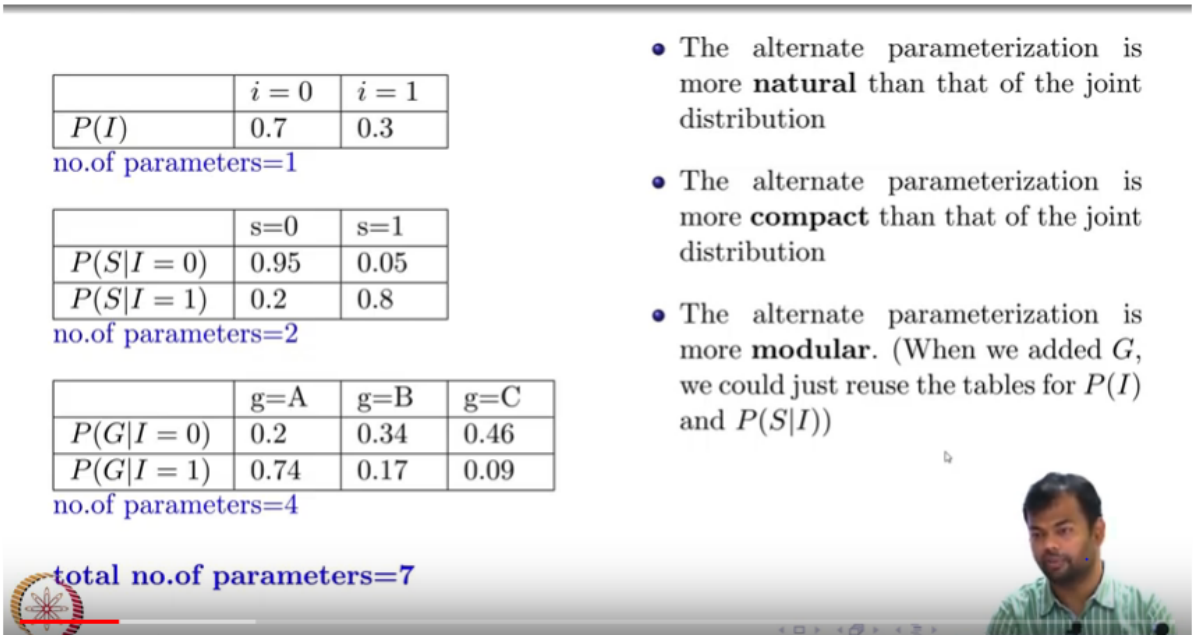
Lecture 16.4

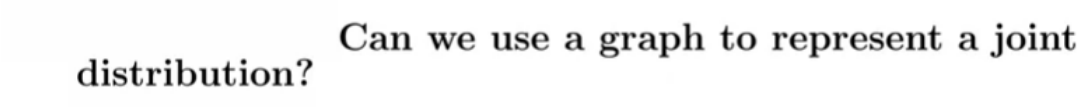
Can we use a graph to represent a joint distribution

Refer slide time :(0:14)



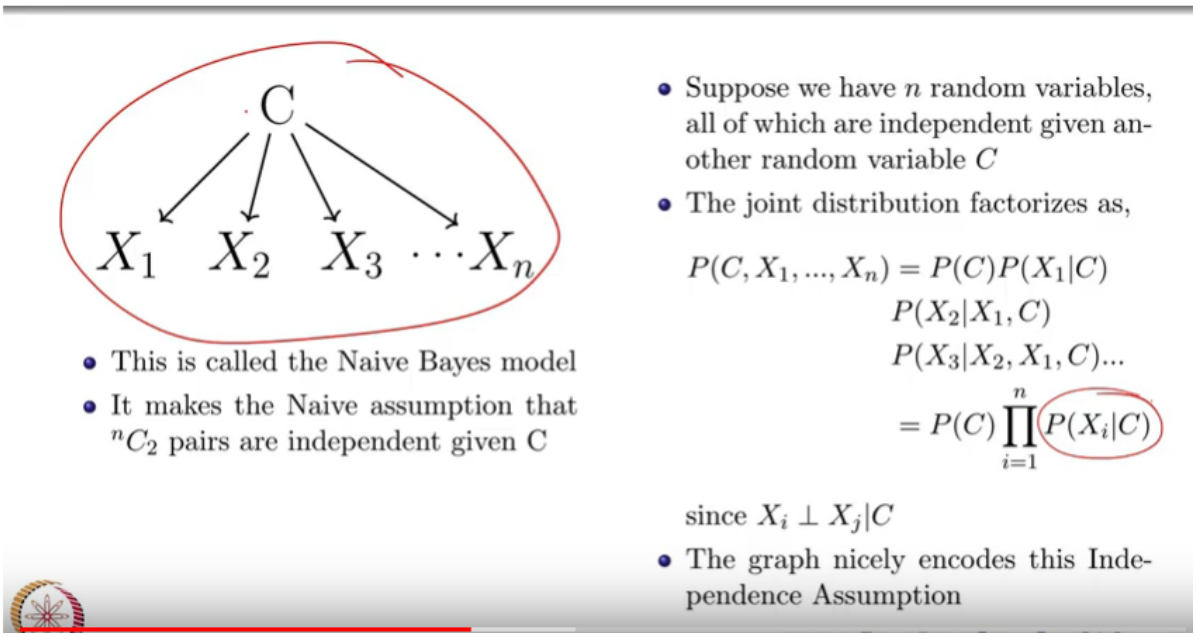
Last class we did a quick recap of probability theory and the main focus was on dealing with joint distributions, along the way we also saw marginal distributions and conditional distributions and the main idea was that if you have a very complex joint distribution involving many random variables, then explicitly representing it would be very difficult because, you end up with to restrain values in the binary case which is very hard to represent. So, we wanted ways of smartly factorizing the joint distribution and we started with the chain rule and then we simplified the chain rule by making certain what assumptions independence assumptions. Right? And that's why in dependences or conditional independencies are very important for the remainder of this section, or the next 20 30 percent of the course because, that will allow us to simplify the joint distribution represented in a more natural compact and modular manner just as we had done for the student case where we had three random variables intelligence, sat score and grades and instead of requiring the 11 parameters that we should have required we were just able to manage with seven parameters of course in this small example it's not really huge ok 11 was also manageable and now you are doing seven but when you only go to many more random variables this benefit would become more obvious. Right? Okay?

Refer slide time :(01:34)



We ended with this question and that's how we will start the next module which is can we use a graph to represent a joint distribution. Right? So, let's see what I mean by that and how to do it

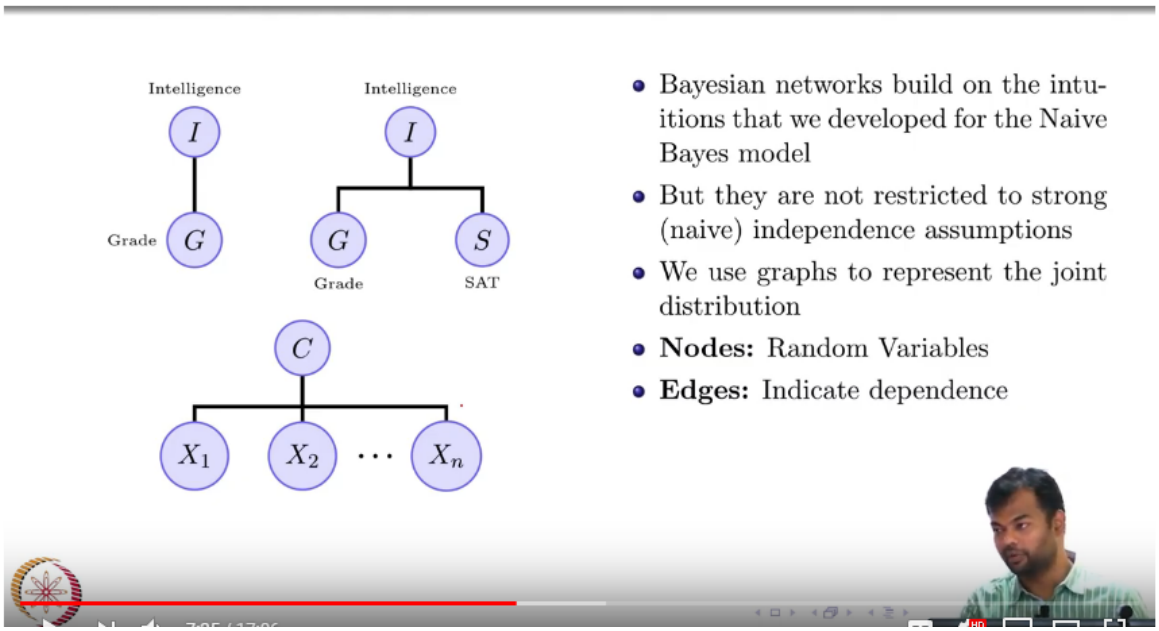
Refer slide time :(01:44)



So, suppose we have n random variables and all of them are independent, given another random variable C ok. So, what is the total number of random variables? That you have n plus 1 Right? Of these n random variables are independent of each other pair wise independent given the random variable C ok. So, now this is one way of representing it Right? Then I am saying that there is a random variable C and this arrow says that all the random variables X1 to Xn actually depend on this random variable C Right? Now how would the Joint Distribution factor is in this case starting with a chain rule what would the joint distribution looks like? how many of you have seen this before please raise your hands high up I'll ask this question again after three minutes and then all of you in raise your hands how would this joint distribution factorize? He'll start with the chain rule. Right? This is what the chain will look like that because, you took C because, and there was a natural ordering. Right? Because, everything depends on C she said C then X 1 given C then X 2 given X 1 C then X 3 given X 2 X 1 C and so, on now what would happen? What are the kind of terms? That you'll, you'll that will remain in this expression Xi given C in general Right? So, this is going to be since all XI's are independent of XJ given C whenever you have C on the condition side you can get rid of all the other variables. Okay? How many of you have seen this before Okay? I'll ask again after one minute and then all of you raise your hands. What model is this everyone? Now is how many of you seen this before all of you. Okay? So, this graph nicely encodes this independence assumption that all the variables are independent of each other given this variable series and that's what the graph is trying to tell you that there are no dependencies between X 1 X 2 and so, on but all of them are dependent on C. Right? And this is actually the naive Bayes model and it makes the assumption that the n choose two pairs of the random variables, are all independent of each other given the random variable C and that's the naive assumption. Right? Because, in practice now let's go back to our famous, now famous

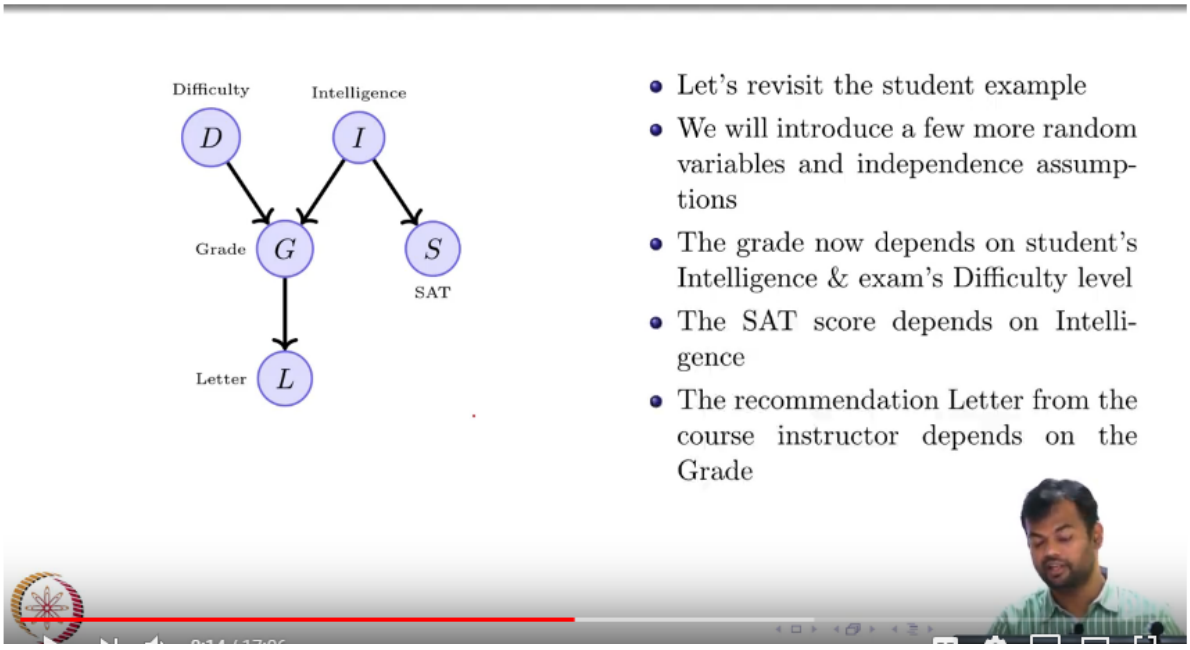
example of drilling oil I mean you don't expect salinity pressure and all that to be independent of each other given say whether the oil exists there or not. Right? But an eye base model actually makes this naive assumption and you will be surprised and in practice, it works quite well it's not a killer model it's not that it's the one solution that you can use for everything but it does reasonably well given the simplicity. Right? And the simplicity now has some repercussions. Right? So, can you go back and tell me that I had stated three problems for the for a general Joint Distribution. Right? What were those three problems? Computational, cognitive, statistical, do you see all three of these actually go away one should make this naive assumption, computational why because the storage requirement is very less, Right? How many tables do you need to store assuming everything is binary interesting, how many tables you need to store for each X I how many tables you need to store one table for every X I Okay? I see why you're saying one I would have said two tables one for the case when C is equal to one and the other for the case when C is equal to zero but Okay? You can put both that information in one single table also that's also fair enough so, you need order n tables. Right? There whether you counted as 2 n plus 1 or n it's still order n. Right? So, that's clearly simplifies now cognitive sure to ask a human that tell me what is the probability of salinity given oil a it's more manageable as compared to asking that person to write down what's the probability of saline is equal to high pressure equal to low and all these things combined as compared to the single thing and statistically even if you don't have a lot of data the number of parameters that you need to estimate is much, much smaller here in fact it's order n again, and those are in the binary case and those are again easy to handle a so, so the key thing here. Right? Which makes the naive Bayes model simple and popular I would say is that making this independence assumption which greatly simplifies the joint distribution. Right? So, this is the idea that we need to latch on to that independence assumptions are important I've been kind of trying to reinforce, this again and again even during the last lecture and the other key idea is that you can use actually a graphs to encode this information. Right? So, graph, graph clearly tells you what the situation is that all the variables are independent of each other given C. Right? So, this is what we are going to build on and naive Bayes model is a very special case of a mission where it's a very special case of a, a Bayesian network

Refer slide time :(06:40)



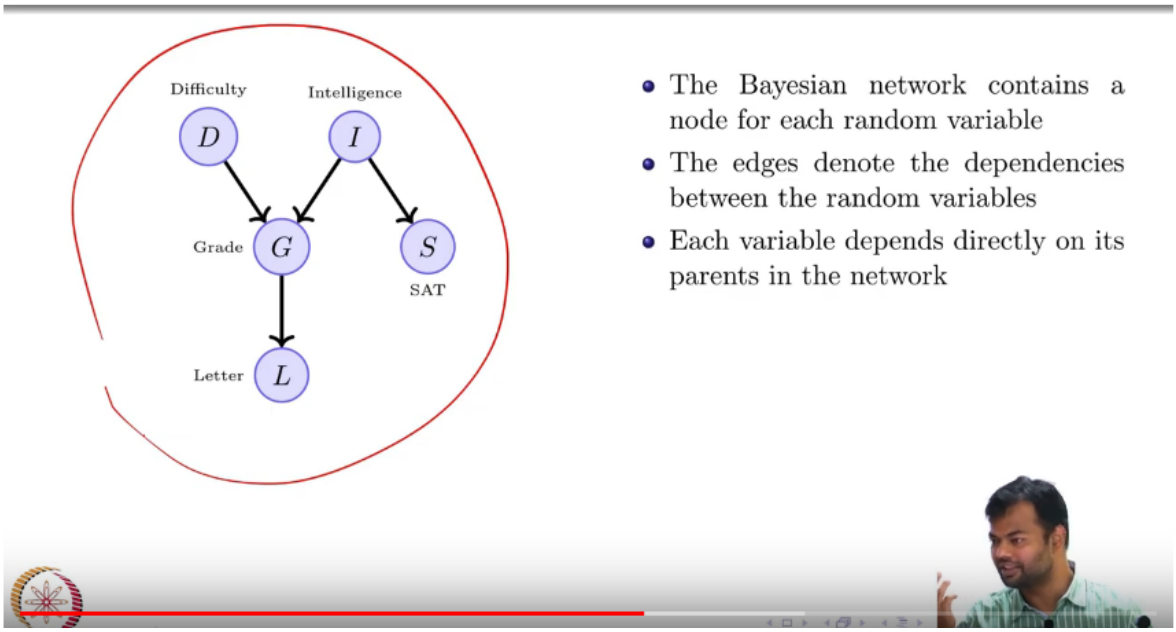
So, Bayesian networks actually build on the intuition that we develop for the naive Bayes model in particular that you can use a graph to represent it and the second thing is that these independence is help you to simplify the Joint Distribution Okay?. So, but unlike the naive based model they are not restricted to these very strong or naive assumptions. Right? I mean there are very strong hence knife. Okay? And you could use graphs to represent joint distributions so, for example the first example that we had where we had only these two random variables intelligence and grade actually we need the arrow here. Okay? So, we have this simple graphical models in that case and the, the, the graph knife since I am talking about graphs you have to tell what are the nodes and what are the edges so, the nodes are random variables and the edges indicate dependence and there are directed edges. So, it tells you which variable depends on which variable it so, there's a semantics encoded in the direction .Okay? So, that's what a Bayesian network or a graphical model or a not a graphical model a directed graphical model or a Bayesian network case. Okay?

Refer slide time :(07:41)



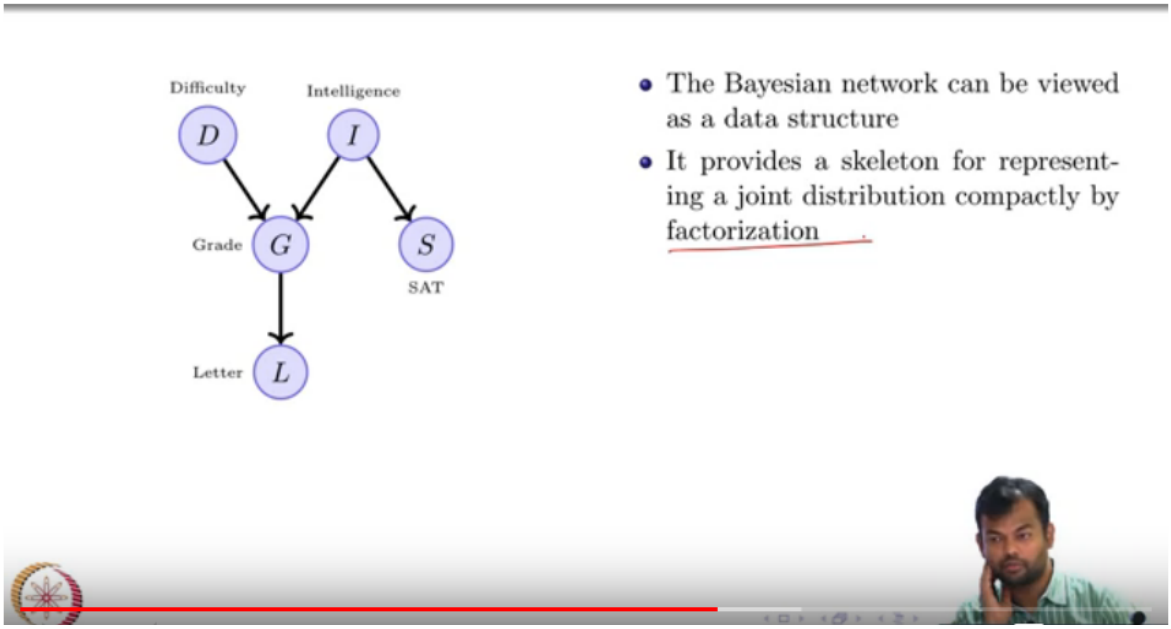
Now let's revisit the student example and we launched a city visited will also introduce a few more random variables and independence of the options. Right? Now can you tell me what are the independence assumptions that I'm making here or at least tell me what are the dependence assumptions That I'm making here what variables depend? On what grade depends, on intelligence and difficulty of the course side score depends, on intelligence and there's a recommendation letter which depends on grade in the course. Right? So, the grade now depends on student's intelligence and the exams difficulty level, the sad score depends on intelligence and the recommendation later depends on the grade. Okay?

Refer slide time :(08:16)



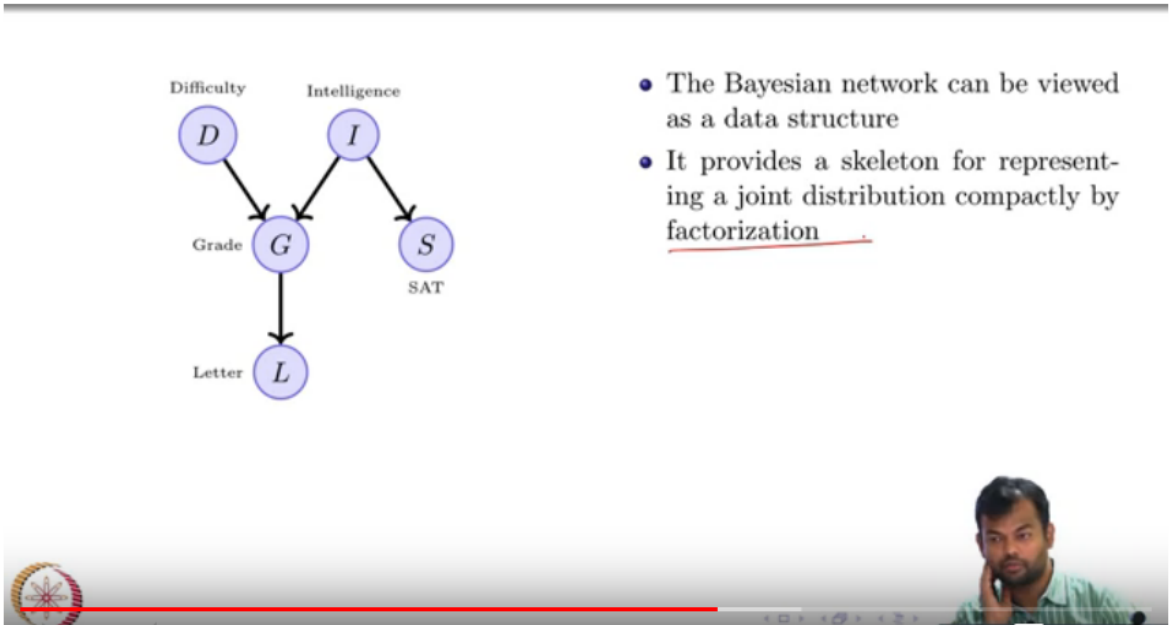
Now the Bayesian network contains a node for each of these random variables and the edges depend denote the dependencies and the direction of the dependency. Okay? Now one thing which is clear is that that's why I asked you tell me what the dependencies are and that's very easy to say because, each variable clearly depends, on its parent and that's how we have actually constructed the graph. Right? So, this graph actually encodes our assumption about how the student world behaves or the simple student world that we are considering it encodes our assumptions, about that is this the only possible graph that I could have drawn, I could have made, a different set, of what is this actually? this is a this is my choice of the model this is how I think the world works or the student world works I could have assumed something different, can you tell me one extra edge that you could have added here the letter could also depend on the sad score I mean typically when someone writes a letter they also ask you your C GPA or other things that's not only the course or the project that you have done and so, and that's also a fair assumption but I made this assumption, or rather the textbook has made this assumption that this is how the student world works and we'll just stick to that and we'll come back to alternate choices also. Okay?

Refer slide time :(09:28)



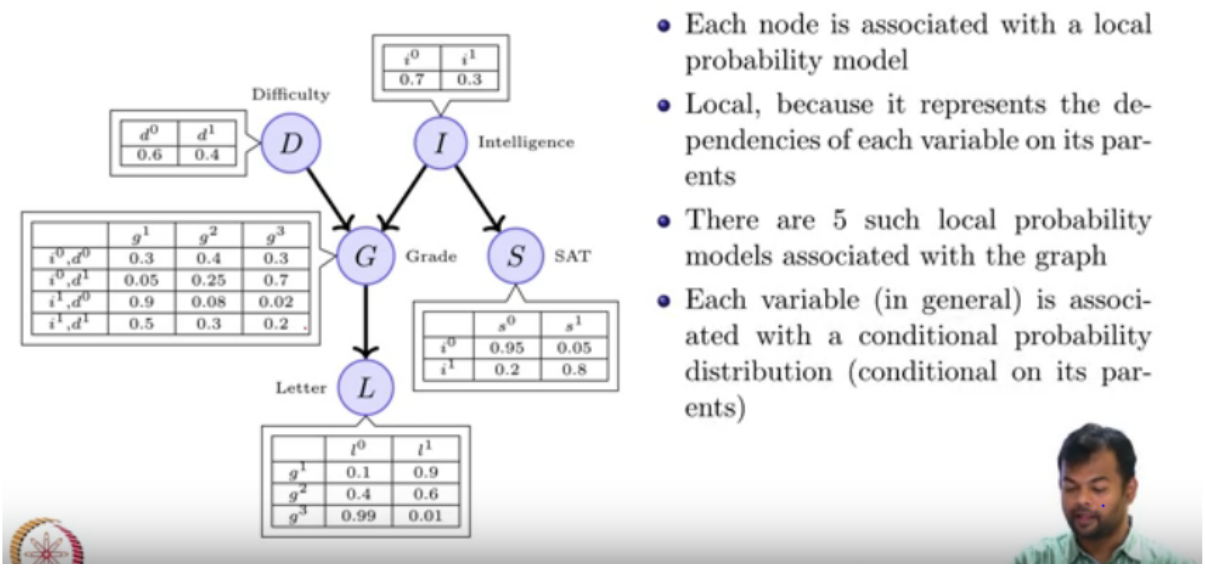
Now the Bayesian network can also be viewed as a data structure, it actually provides a skeleton for representing a joint distribution compactly by factorization. What do I mean by that actually? How can you use the setter as a data structure? What are you going to store in the graph? For every node what do you think you should need to store the graph gives you a factorization. Right? So, for every node can you tell me what will you store along with that node definitely not the joint product distribution? Okay? Then what are the other two options conditional and marginal. So, which node here I will have a marginal distribution associated with it I n D because, they don't depend on anything which are the nodes which would have a conditional distribution associated, with them and you see that if you have all this information you can actually compute the joint distribution. Right? So, this graph is a very neat way of looking at it as a data structure, where you have these nodes and each node has the local probability distributions and from all these local probability distributions, you can compute the joint distribution. Why do I call it local?

Refer slide time :(10:36)



Why do I call it local? Why locally yeah it just talks about its parents and it only says this table a actually only has this it does not have all the variables. Right? That's why it's not global? that's very easy to say it does not have s and L in particular it only has the local neighbourhood that matters at that point right it only has the variable of interest and its immediate parents, it does not have anything else so, it's definitely not global hence locally is that fine and there are five such local probability distributions that you have but this is not what I wanted I wanted the joint distribution why did I what's the point of having these local priority distributions, is not going to give me anything I want the joint distribution what do I do with this local distributions what can you do? What does the graph actually give you A dash of the Joint Distribution a factorization Right? So, the graph is actually giving you all the factors that you need for the joint distribution Right?

Refer slide time :(11:35)



And what if I write the joint distribution or we have already simplified it Okay? So, what we should have done is actually probably written the chain rule. Right? And based on the dependencies encoded in the graph, we could have got rid of certain variables in the chain rule and this is what we would have left fit how many distributions are there on the right hand side five and where are these distributions located they are on the graph. Right? So, these five distributions no associated with the relevant nodes give us all the information that we need to compute the joint distribution. Right? So, this gives a very natural way of both visualizing it as well as storing the information that is relevant and now in this five variable case tell me how many parameters do you have now? Remember that four of these variables are binary and one of them is takes three values. So, what is the total number of variables that you need for the joint distribution? For a binary one is no, no I how many would you need for a joint distribution, on factor is joint distribution if I want to give you the full table. How many values? Do I need to give you how me 48 is that fine 2 into 2 into 2 into 2 into 3. Right? How many do I have here? 15 is that all how many did you count here why good. So, because we have to remove that summation equal to 1 right we have to account for that so, 1 here 1 here 8 here 1 here and 3 here. Right? Is that fine with twin – oh this one has 2 yeah hey so, we are we have a total of 50. Right? So, now you see even with 5 variables you do see a significant reduction. Right? Now it's not just 7 vs. 11 but 15 versus 48 like one-fourth or one-third of what we needed earlier and as the number of variables grow and has a number of possible values that these variables can take grow you'll see much more reduction once you start making these independence assumptions. Right? Okay? And from this joint distribution or from this conditional distribution so, this is known as a conditional parameterization of the Joint Distribution because, your parameters are all conditionals or marginal's and from this conditioners and marginal's you can actually compute any joint distribution that you want. Right? So, in particular if I want you to compute this how will you do it so you want P of I equal to 1 so we'll take this value P of D equal to 0 is the next thing and so on. Right? So, P equal P of G is equal to B that means the second possible grade when I is 1 and D is equal to 0. So, which entry are you going to take Right? Avian gets this I mean it's very straightforward but I just want to go over it once. Right? And you can convince yourself that any joint distribution that I asked you or any particular configuration that I asked you for these five variables from these tables you can compute

them. Right? Okay? So, that's what this entire structure which means the graph which includes the nodes and the edges along with all the conditional of the local probability distributions that you see with each of these nodes is together known as the Bayesian network. Okay? Right? So, now this section it is again to motivate, these are handcrafted this is your assumption of see how the world works. Right? I'm based on this assumption now you could think of it as that just as when I made the assumption suppose, Y and X are scalars. Okay? If, I made the assumption that Y is equal to MX plus C. Right? Then basically I've made the assumption that I only need to learn two parameters for this as opposed to if I made the assumption that Y is equal to w25 into X raised to 25 the polynomial that we had seen, then I had made the assumption that we need to learn 25 parameters for this but that's effectively how it translates to Right? I did not make an assumption that I need to learn 25 parameters I made a Ramson for the model but that's how it translate to the number of parameters, here we have made an assumption that this is how the world works and that has repercussions in terms of number of parameters that you learn and that has applications in there are terms of amount of data you need and so, on if you add an extra edge here or there the number of parameters will increase. Right? Or if you make some more dependence independence assumption said if you just say that I'll ditch difficulty it just depends on intelligence and the number of parameters will change and so, on it so that's completely your modelling choice. Right? So, that is very important to understand when you do machine learning that you come up with the model after that perhaps everything is automated and learning, learning from the data is automated but you have to tell me what the model is and depending on that everything will change. Right? So, that's I cannot emphasize enough that this is the choice the modelling choice that you have made. Okay? And please appreciate that this ties back to what we have been saying in terms of data model parameters and so on. Right? We always considered some models you, decided to use a recurrent neural network even though I had it on the slides but you decided to use it for language modelling or whatever it and that's why you wrote those complex equations and the number of parameters and so on. Okay?